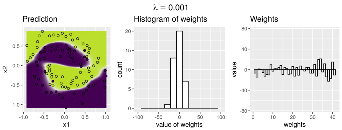
Introduction to Machine Learning

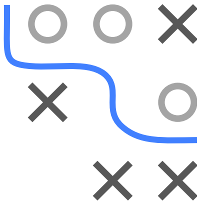
Regularization in Non-Linear Models and Structural Risk Minimization



Learning goals

- Understand that regularization and parameter shrinkage can be applied to non-linear models
- Know structural risk minimization

SUMMARY: REGULARIZED RISK MINIMIZATION



If we should define (supervised) ML in only one line, this might be it:

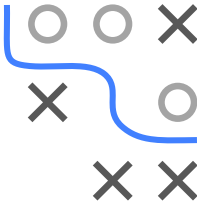
$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \lambda \cdot J(\boldsymbol{\theta}) \right)$$

We can choose for a task at hand:

- the **hypothesis space** of f, which determines how features can influence the predicted y
- the **loss** function L, which measures how errors should be treated
- the **regularization** $J(\boldsymbol{\theta})$, which encodes our inductive bias and preference for certain simpler models

By varying these choices one can construct a huge number of different ML models. Many ML models follow this construction principle or can be interpreted through the lens of regularized risk minimization.

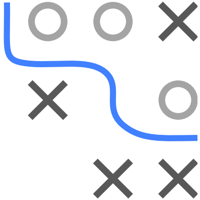
REGULARIZATION IN NONLINEAR MODELS

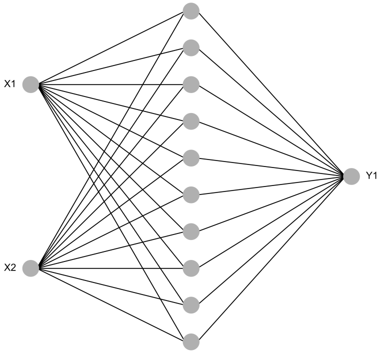


- So far we have mainly considered regularization in LMs.
- Can also be applied to non-linear models (with numeric parameters), where it is often important to prevent overfitting.
- Often, non-linear models can be seen as LMs based on internally transformed features.
- Here, we typically use $L2$ regularization, which still results in parameter shrinkage and weight decay.
- Adding regularization is commonplace and sometimes crucial in non-linear methods such as NNs, SVMs, or boosting.
- By adding regularization, prediction surfaces in regression and classification become smoother.

REGULARIZATION IN NONLINEAR MODELS

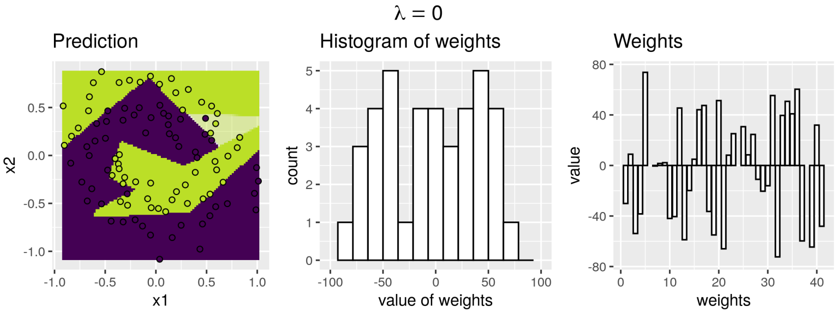
Setting: Classification for the `spirals` data. Neural network with single hidden layer containing 10 neurons, regularized with $L2$.
Varying λ affects smoothness of the decision boundary and magnitude of network weights:





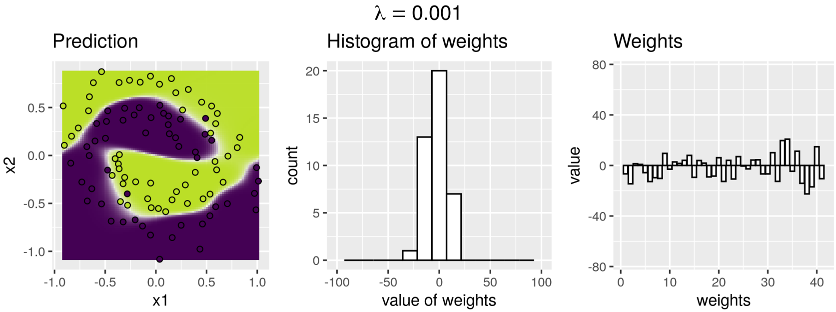
REGULARIZATION IN NONLINEAR MODELS

Setting: Classification for the `spirals` data. Neural network with single hidden layer containing 10 neurons, regularized with $L2$.
Varying λ affects smoothness of the decision boundary and magnitude of network weights:



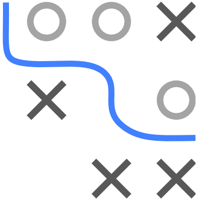
REGULARIZATION IN NONLINEAR MODELS

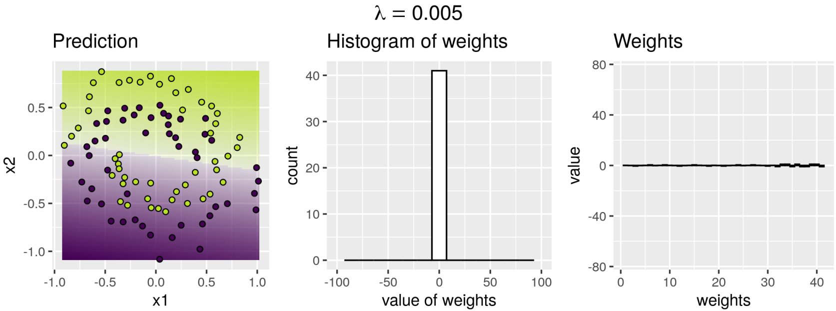
Setting: Classification for the `spirals` data. Neural network with single hidden layer containing 10 neurons, regularized with *L*2.
Varying λ affects smoothness of the decision boundary and magnitude of network weights:



REGULARIZATION IN NONLINEAR MODELS

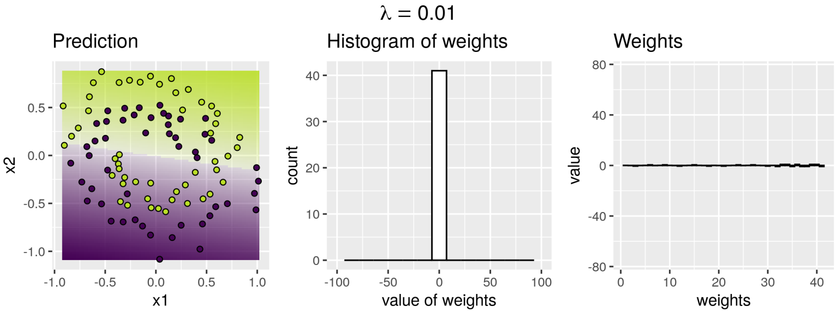
Setting: Classification for the `spirals` data. Neural network with single hidden layer containing 10 neurons, regularized with *L*2.
Varying λ affects smoothness of the decision boundary and magnitude of network weights:



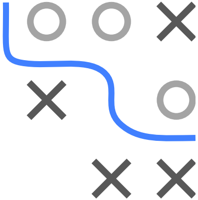


REGULARIZATION IN NONLINEAR MODELS

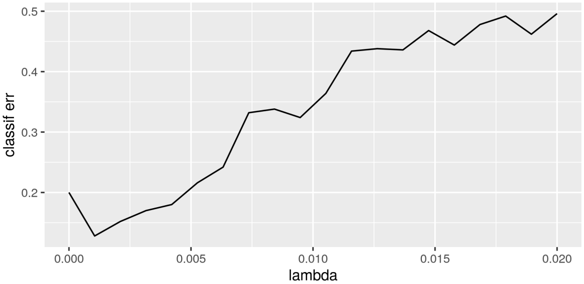
Setting: Classification for the `spirals` data. Neural network with single hidden layer containing 10 neurons, regularized with *L*2.
Varying λ affects smoothness of the decision boundary and magnitude of network weights:



REGULARIZATION IN NONLINEAR MODELS



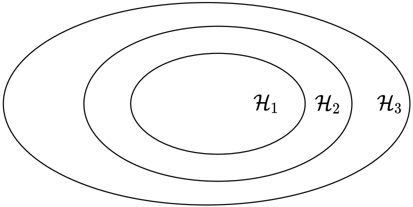
The prevention of overfitting can also be seen in CV. Same settings as before, but each λ is evaluated with repeated CV (10 folds, 5 reps).



We see the typical U-shape with the sweet spot between overfitting (LHS, low λ) and underfitting (RHS, high λ) in the middle.

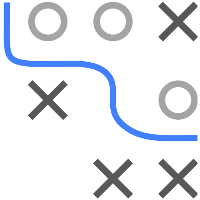
STRUCTURAL RISK MINIMIZATION

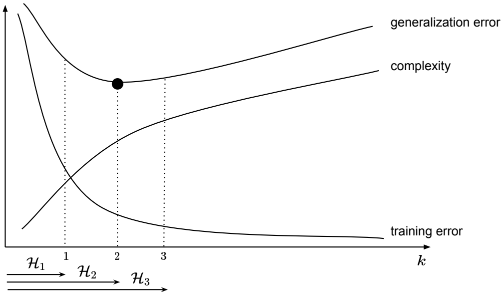
- Thus far, we only considered adding a complexity penalty to empirical risk minimization.
- Instead, structural risk minimization (SRM) assumes that the hypothesis space $\mathcal{H}$ can be decomposed into increasingly complex hypotheses (size or capacity): $\mathcal{H} = \cup_{k \geq 1} \mathcal{H}_k$.
- Complexity parameters can be the, e.g. the degree of polynomials in linear models or the size of hidden layers in neural networks.



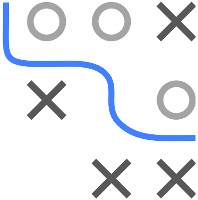
STRUCTURAL RISK MINIMIZATION

- SRM chooses the smallest k such that the optimal model from $\mathcal{H}_k$ found by ERM or RRM cannot significantly be outperformed by a model from a $\mathcal{H}_m$ with $m > k$.
- By this, the simplest model can be chosen, which minimizes the generalization bound.
- One challenge might be choosing an adequate complexity measure, as for some models, multiple complexity measures exist.



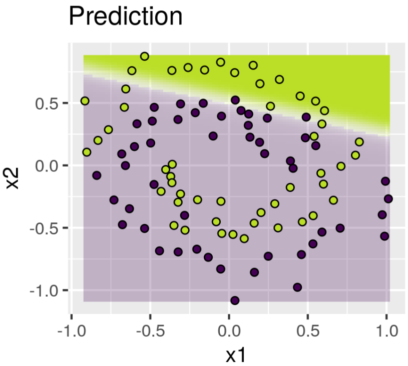


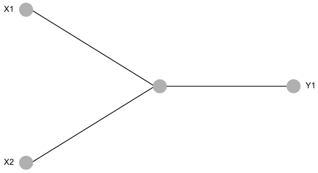
STRUCTURAL RISK MINIMIZATION



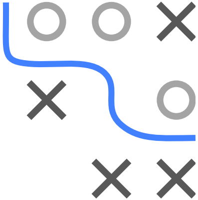
Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:

size of hidden layer = 1

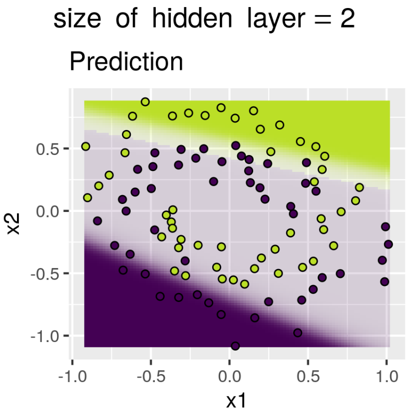


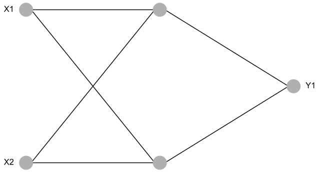


STRUCTURAL RISK MINIMIZATION



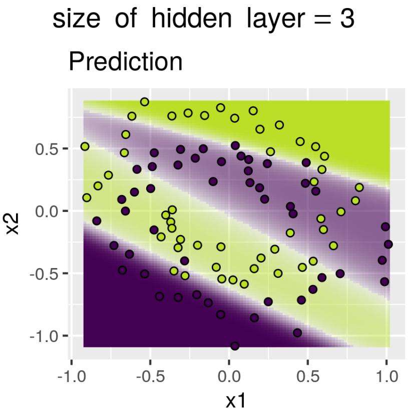
Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:

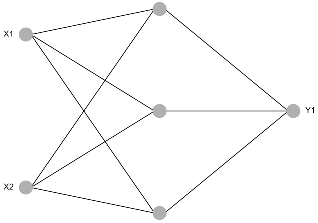




STRUCTURAL RISK MINIMIZATION

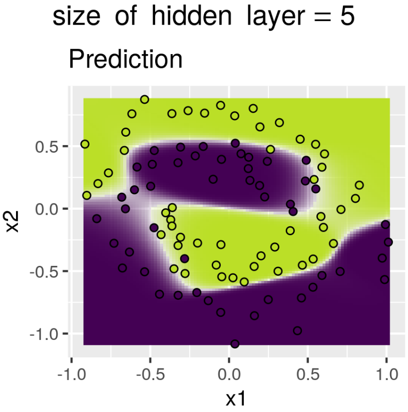
Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:

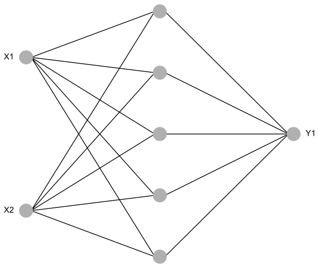




STRUCTURAL RISK MINIMIZATION

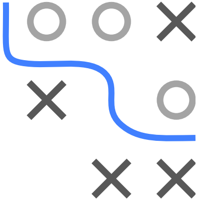
Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:

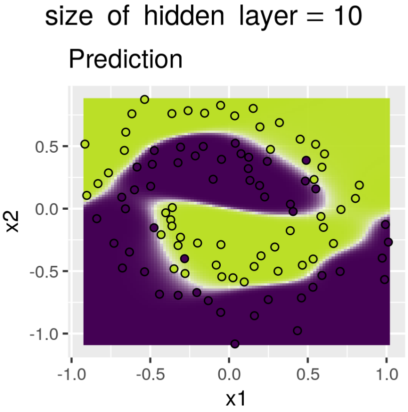


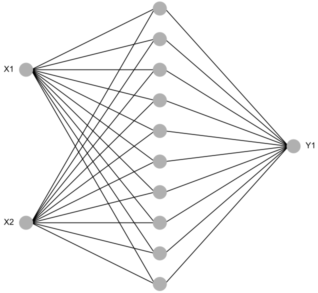


STRUCTURAL RISK MINIMIZATION

Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:



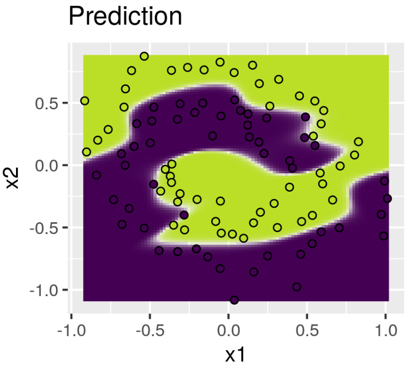


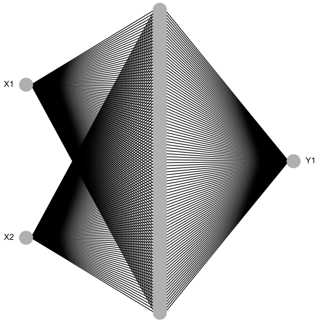


STRUCTURAL RISK MINIMIZATION

Setting: Classification for the `spirals` data. NN with 1 hidden layer, and fixed (small) L2 penalty.
Varying the size of the hidden layer affects smoothness of the decision boundary:

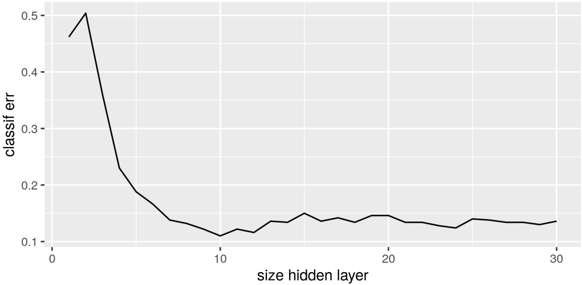
size of hidden layer = 100





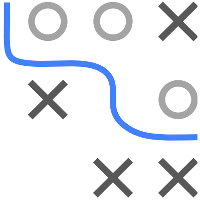
STRUCTURAL RISK MINIMIZATION

Again, complexity vs CV score.



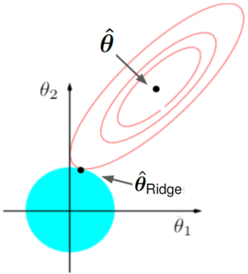
A minimal model with good generalization seems to have ca. 6-8 hidden neurons.

STRUCTURAL RISK MINIMIZATION AND RRM



Note that normal RRM can also be interpreted through SRM, if we rewrite the penalized ERM as constrained ERM.

$$
\begin{aligned}
\min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) \\
\text{s.t.} \quad & \|\boldsymbol{\theta}\|_2^2 \leq t
\end{aligned}
$$



We can interpret going through λ from large to small as through t from small to large. This constructs a series of ERM problems with hypothesis spaces $\mathcal{H}_\lambda$, where we constrain the norm of $\boldsymbol{\theta}$ to unit balls of growing size.